CAR POPULARITY PREDICTION: A MACHINE LEARNING APPROACH

[1]T. RAVI KIRAN KUMAR, [2]N. AKASH GOUD, [3]S. PRAVEEN GANDHI,[4]M. POOJA,[5]AYESHA BEGUM

[1]ASSISTANT PROFESSOR, [2,3,4&5] UG STUDENTS

DEPARTMENT OF CSE, MNR COLLEGE OF ENGG. & TECHNOLOGY, MNR NAGAR, FASALWADIGUDA, SANGA REDDY-502294

ABSTRACT

Today is a world of technology with a foreseen future of a machine reacting and thinking same as human. In this process of emerging Artificial Intelligence, Machine Learning, Knowledge Engineering, Deep Learning plays an essential role. In this paper, the problem is identified as regression or classification problem and here we have solved a real-world problem of popularity prediction of a car company using machine learning approaches.

INTRODUCTION

In the era which we live in, technology has a big impact on our lives. Artificial intelligence [6], knowledge engineering, Machine learning, Deep learning [4][5], Natural language processing[7][8] are emerging technologies which plays an important role in the leading projects of today's world. Artificial intelligence is an area or branch which aims or emphasizes on creating machine that works intelligently and their reactions is similar to that of human. In Artificial Intelligence, Machine learning is an essential and core part providing the ability of learning and improving by itself. The focus of this technique is on creation of programs which can pick the data and learn from it by itself. Earlier, statistician and developers worked together for predicting success, failure, future etc. of any product. This process led to delay of the product development and launch. Maintenance of such product in the changing technology and data is also one of the major challenges. Machine learning made this process easier and faster. There are various Machine learning algorithms broadly categorized into four paradigms: • Supervised learning [7] [9] [10]: This learning algorithm provides a function so as to make predictions for output values, where process starts from analysis of a known training dataset. This algorithm can be applied to the past learned data to new data using labels so as to predict future events. • Unsupervised learning: This algorithm is used on training dataset and informs which is neither classified nor labeled. It also studies to infer a function from a system to describe a hidden structure from unlabeled data. Clustering is an approach of unsupervised learning. • Semi supervised learning [6] [11]: It takes the characteristics of both unsupervised learning and supervised learning. These algorithms uses small amount of labeled data and large amount of unlabeled data. • Reinforcement [12]: In this algorithm, interaction is made to environment by actions and discovering errors. It allows machines and software agents in determining ideal behavior in a specific context such that performance could be maximized. Regression and Classification problems are types of problems in supervised learning. In classification, conclusion is drawn using values which are obtained by observation. A discrete output variable say y is approximated by this problem using a mapping function say f on input variables say x. The output of classification is generally discrete but it can also be continuous for every class label in the form of probability. A regression problem has output variable as a real or continuous value. A continuous output variable say y is approximated by this problem using a mapping function say f on input variables say x. The output of regression is generally continuous but it can also be discrete for any class label in the form of an integer. A problem with many output variables is referred to multivariate regression problem. In this paper we will be focusing on a problem picked from hackerrank where a company is trying to launch a new car modified on the basis of the popular features of their existing cars. The popularity will be predicted using machine learning approach. It can be classified as regression problem especially a multivariate regression problem and the problem can be classified under supervised learning. Thus various supervised learning algorithms will be used for this prediction.

LITERATURE SURVEY

In paper "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks[1]", author has reviewed some classification algorithms such as random forest, gradient boosted trees, artificial neural network and logistic regression to predict 463 stocks of the S&P 500. In order to study the predictability of these stocks, author has performed multiples of experiments with these classification algorithms. The obtained result of predicting future prices from the past available data was not up to the mark as the expected result, The author wanted to obtain. However, they successfully showed the vast growth in predictability of European and Asian indexes closed a little while back.

In paper "Performance evaluation of predictive models for missing data imputation in weather data[2]", author has suggested a new approach to manage the missing data in weather data by performing various tests with NCDC dataset to assess the prediction error of five methods: linear regression, SVM, random forest, KNN Implementation and kernel ridge. In order to handle the missing values of dataset they performed two actions: 1.removing the entire row which contains missing value and 2. Impute the missing data. They performed both the methods to handle the missing data and compared the observed result. In paper "Amazon EC2 Spot Price Prediction using Regression Random Forests [3]", author has proposed Regression Random Forests (RRFs) model to forecast the Amazon EC2 Spot Price one week ahead and one month ahead. This prediction model would help in planning when to acquire the spot instance, the model also predicts the execution cost and it also suggests the user when to bid in order to minimize the execution cost.

IN "JIAO, YANG, AND JÉRÉMIE JAKUBOWICZ. "PREDICTING STOCK MOVEMENT DIRECTION WITH MACHINE LEARNING: AN EXTENSIVE STUDY ON S&P 500 STOCKS." BIG DATA (BIG DATA), 2017 IEEE INTERNATIONAL CONFERENCE ON. IEEE, 2017."

Stocks movement direction forecasting has received a lot of attention. Indeed, being able to make accurate forecasts has strong implications on trading strategies. Surprisingly enough little has been published, relatively to the importance of the topic. In this paper, we reviewed how well four classic classification algorithms: random forest, gradient boosted trees, artificial neural network and logistic regression perform in predicting 463 stocks

of the S&P 500. Several experiments were conduced to thoroughly study the predictability of these stocks. To validate each prediction algorithm, three schemes we compared: standard cross validation, sequential validation and single validation. As expected, we were not able to predict stocks future prices from their past. However, unexpectedly, we were able to show that taking into account recent information – such as recently closed European and Asian indexes – to predict S&P 500 can lead to a vast increase in predictability. Moreover, we also found out that, among various sectors, financial sector stocks are comparatively more easy to predict than others. Stock market has long been characterized by its dynamic, complicated, and non-stationary nature [1]. Market movements are dependent upon various factors ranging from political events, firms policies, economic background, commodity prices, exchange rates, movements of other stock markets to psychology of investors [2], [3]. In addition, the Efficient Market Hypothesis [4] assumes that asset prices are fair and adjust quickly to reflect all past and present information, which implies that future stock price movements are independent from pending and past information and should therefore follow a random walk pattern. If this hypothesis were true, then any attempts to predict the market would be fruitless [5]. The EMH hypothesis has been tested extensively across various markets. The results are, however, sometimes contradictory. Many early work support the random walk model [6]. “There is no other proposition in economics which has more solid empirical evidence supporting it than the Efficient Market Hypothesis”, as said by Jensen [7]. However, modern studies [8], [9] on stock markets reject the random walk behavior of stock prices. Besides the efficient market hypothesis, there are two schools of thought regarding stock market predictions: fundamental analysis and technical analysis. Fundamental analysis [10] consists of evaluating the intrinsic value of a stock by examining the financial condition of a company. However, the proponents of the EMH argue that the intrinsic value of a stock is always equal to its current price. Technical analysis, on the other hand, is a study of the market itself. Technical analysts believe market action tells everything, so price and trading volume time series are enough for prediction tasks. Since market driving forces (i.e., human psychologies) hardly change, the prices are then considered to be recurrent and predictable since history always repeats itself. In this paper, we took the technical analysis viewpoint and tried to predict stock market movements using historical stock prices and modern tools from machine learning and artificial intelligence. In other terms, we asked the following question: to what extent are market stock prices self predictable? Technical analysts traditionally build compound features from historical data, called technical indicators, representing various aspects of a stock in order to exploit recurring patterns. Some commonly seen technical indicators include MA (moving average), RSI (Relative Strength Index), MACD (Moving Average Convergence/Divergence Oscillator), CCI (Commodity channel index) etc. In our study we regard the movement of each stock price as a time series and perform extensive feature extraction to obtain over 200 features for each stock on a given time window. Similar to traditional technical indicators, our newly extracted features aim to capture different aspects of a stock thus reveal potential predictive power to stock movement. We studied 463 stocks, which are constituents of S&P 500 index, with over 7 years of trading history. We examined several classification models: logistic regression, artificial neural network, random forest, and gradient boosted trees to predict the direction of tomorrow based on the information of today. We also evaluated the usefulness of 8 global market index, including 3 Asian index (Nikkei 225, Hang Seng, and All Ords), 2 Europe index (DAX, FTSE 100) and 3 US index (NYSE Composite, Dow Jones Industrial Average, S&P 500). It is worth noticing that Asian and Europe Markets close before US markets, therefore they can be used to provide additional information to predict stocks of US markets. In our numerical experiments, we first performed detailed feature selection revealing features with the most predictive power. Next we fine tuned 4 state of the art classification algorithms and compared their prediction performance on all of 463 stocks. Then we compared three different validation schemes for model selection and parameter tuning and confirmed the usefulness of time-aware cross-validation. Further on, we analyzed the predictability of stocks within different sectors and compared the prediction performance on stocks per sector, which could provide useful advise on stock investment. Last, we compared models with the one proposed in a recent and popular study on predicting S&P 500 index movements direction and verified the efficiency of our models. In summary, our main contributions are the following: 1) The scope of our study is unprecedented in the existing literature, to the best of our knowledge. With 463 stocks and 8 index across the globe being analyzed, more than 200 technical indicators used as features, 4 state-of-the-art classification models involved, we conduct an extensive analysis and comparison of different prediction approaches. 2) We provide a publicly available notebook to make our study easily reproducible 1. The data used in this paper is also provided in the project folder. 3) We highlight that data and feature selection play a key role in such prediction task whereas prediction performance improvement due to fine tuning remain insignificant. 4) We demonstrate that, for stock market price movement prediction, immediate past contains most of the signal. 5) We find that stocks within financial sectors are the most predictable ones with more than 10 point of prediction accuracy above the overall average In recent years, there have been a growing number of studies looking at the direction of movements of various kinds of financial instruments. Both academia and practitioners have made tremendous efforts to predict future movements of stock market prices and devise financial trading strategies to translate forecasts into profits [11]. The emergence of machine learning and artificial intelligence algorithms has made it possible to tackle computationally demanding models for stock price movement direction prediction. In [12], the author shows that AI outperformed traditional statistical methods in dealing with various financial problems including credit evaluation, portfolio management and financial prediction/planning. In our study, we are interested in forecasting the direction of stock price movement. Naturally, an associated trading strategy takes a short position when direction is predicted to go down and a long position when the predicted direction is up. In [13], the author used 53 technical indicators to predict direction of three stocks and one index of Taiwan stock market with a SVM based approach and a prediction accuracy between 55% and 65%. In [14], the author tried to predict Dow Jones Industrial Average index using three models including DT, KNN, and NN. After selecting the most predicable period using Hurst exponent, the author restrained the scope of the study to this period and then performed a voting based ensemble methods to combine the result of those three models to achieve a better accuracy than any single classifier. In [15], the author used 10 technical indicators to predict Istanbul Stock Exchange National 100 Index (ISE) and reported an over 75% accuracy using neural network. In [16], the author tried to predict with 10 technical indicators the direction of stock movement for Indian stock market by using two stocks and two index as samples. Attempts of predicting stock movement without the use of technical indicators have also been made recently. In [17], the author tried to predict Korean and Hong Kong market using price data alone with SVM based approach preceded by a PCA to reduce the

dimension of input features. In [18], the authors used Random Forest to predict stock direction of three stocks: Apple, Samsung and GE. A recent survey [19] compares 11 classification algorithms on 71 different datasets and shows that Gradient Boosted Decision Tree (GBDT), followed by Support Vector Machine (SVM) and Random Forest (RF) are the most accurate among their competitors. It is also worth noticing that GBDT are rarely used in previous studies on stock direction prediction however the model shows its superior prediction capacity against other models in many recent Kaggle competitions 2. Therefore, in our study, we include GBDT into our candidate list along with other most commonly used methods such as RF, ANN and Logistic Regression. SVM is excluded in our study because of its lack of capacity to naturally provide probability estimation of its prediction

IN "GAD, IBRAHIM, AND B. R. MANJUNATHA. "PERFORMANCE EVALUATION OF PREDICTIVE MODELS FOR MISSING DATA IMPUTATION IN WEATHER DATA." ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS (ICACCI), 2017 INTERNATIONAL CONFERENCE ON. IEEE, 2017"

Real datasets can have missing values for a different reasons such as in data that were not kept on file and data corruption. Climate forecasting has a highly relevant effect in agricultural fields and industries sectors. The process of predicting climate conditions is required for different areas of life sectors. Handling missing data is significant because a lot of machine learning algorithms performance are affected by missing values in addition, they do not support data with missing values. Various techniques have been used to process missing data problem and the most applied is removing any row that contains at least one missing value. Also, another approaches to solve missing data problems are to impute the missing data to yield a more complete dataset. In order to improve the accuracy of prediction with the climate data, missing value from dataset should be removed or imputed/predicted in the pre-processing phase before using the data for prediction or clustering in the analysis step. In this paper, we propose a new technique to handle missing values in weather data using machine learning algorithms by execute experiments with NCDC dataset to evaluate the prediction error of five methods namely the kernel ridge, linear regression, random forest, SVM imputation and KNN imputation procedure. The missing values were imputed using each method and compared to the observed value. Results of the proposed method were compared with existing techniques. The most common problem in real dataset and statistical analysis is missing data. The percentage of missing values varies from one dataset to another. Generally, the dataset contains different percentages of missing values in each column [20]. Usually, if the rates of missing data is less than 1% are called trivial and missing data ratio with in the range of 1−5% is flexible. The advanced methods are applied to handle rates with in the range 5 − 15%, and greater than 15% have a very great impact on analysis [4, 12]. The process of missing data estimation is an essential problem in data analysis , and there are several solutions have been suggested to solve such problem , for instance, statistics and data mining [16, 17]. The most familiar approaches to handle missing values involves removing all instances with missing values from a dataset and imputing missing values [22, 25]. Imputation is a method to assess the missing data based on the complete instances in a dataset. The types of imputation methods are parametric and nonparametric regression imputation methods [11]. In the imputation strategy such as data mining and machine learning algorithms, missing data handling is independent of data analysis algorithms. Usually, the observed data in incomplete rows is used to estimate missing values by imputation algorithms. For example, KNN imputation algorithm depends on the available data to indicate neighbours of an instance with missing values , and the class of the instance in clustering-based imputation algorithms [8, 24]. There are three categories of treatment methods for missing data namely :- a) case deletion, which is the most common used. So that each row containing missing values is deleted. b) parameter estimation can be done by maximum likelihood procedures that are used to treat parameter estimation. In general, the methods of parameter estimation are more efficient than case deletion methods, because they can use all the data available in the dataset. However, parameter estimation suffers from the following limitations: a high sensitivity to outliers, and a high degree of complexity and c) imputation techniques: the definition of imputation is a procedure to fill missing data with predicted ones based on values available in the dataset [6, 10]. In this paper the following techniques are selected to fill missing data: the kernel ridge , linear regression, random forest, SVM imputation and the k-nearest neighbor (KNN) imputation procedure. In a NCDC dataset, the attributes that have missing values are related with temperature attribute then by using known values of temperature we can predict the unknown values of the other attributes. For evaluating the results, the performance metrics considered are standard deviation of error (STDE), variance of error (VARE), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), bias and coefficients of determination R2 . Swati et all [10] proposed a methodology consists of two phases. The first phase is missing value check and outlier detection, this step is pre-processed step of missing data and missing value locations are checked in the input dataset. The second step is calculating estimation of missing values, firstly small array is created from the input data in which missing data value is existing. secondly, calculate centroid of the subset, centroid is generated by the mean of subset. The value of Xest(estimated value) is separately computed for every missing value in the complete dataset. Gimpy et all [21] proposed estimation of missing values using decision tree approach. They explored that the performance of classifier affected by the existence of missing values in a dataset. Dataset that used in this study contains some missing values, which consists of student data of university system. Classification algorithm (C4.5/J48) is used to fill missing values and the accuracy is measured by confusion matrix. Engels et all [9] suggested a fourteen types of mean imputation techniques introduced on missing data(Column mean - Column median - Class mean - Class median - Hot deck - Regression - Regression with error - Previous row mean - Previous row median - Last observation carried forward - Row mean - Row median - Next observation carried backward - Average of the last known and next known values). The missing value was imputed using each method and compared to the observed value. Methods were compared on the root mean square error, bias, mean absolute deviation and relative variance of the estimates. Sallam et all [18] suggested handling numerical missing values using rough sets. They investigate multiple ways used to solve the problem of missing values in a a dataset. The proposed model used rough set theory as a technique to fill missing data. In addition, the model has ability to estimate the missing values for condition and decision attributes. Patidar et all [15] proposed handling missing value using decision tree algorithms. They apply data mining techniques to analysis the performance of the students in educational system. C5.0 decision tree is used to make analysis and making decisions. Also, comparison is accomplished by ID3, C4.5 and C5.0. This research explored that C5.0 gives more accurate and efficient output than the other methods.

IN "KHANDELWAL, VEENA, ANAND CHATURVEDI, AND CHANDRA PRAKASH GUPTA. "AMAZON EC2 SPOT PRICE PREDICTION USING REGRESSION RANDOM FORESTS." IEEE TRANSACTIONS ON CLOUD COMPUTING, 2017."

Spot instances were introduced by Amazon EC2 in December 2009 to sell its spare capacity through auction based market mechanism. Despite its extremely low prices, cloud spot market has low utilization. Spot pricing being dynamic, spot instances are prone to out-of bid failure. Bidding complexity is another reason why users today still fear using spot instances. This work aims to present Regression Random Forests (RRFs) model to predict one-week-ahead and one-day-ahead spot prices. The prediction would assist cloud users to plan in advance when to acquire spot instances, estimate execution costs, and also assist them in bid decision making to minimize execution costs and out-of-bid failure probability. Simulations with 12 months real Amazon EC2 spot history traces to forecast future spot prices show the effectiveness of the proposed technique. Comparison of RRFs based spot price forecasts with existing non-parametric machine learning models reveal that RRFs based forecast accuracy outperforms other models. We measure predictive accuracy using MAPE, MCPE, OOBError and speed. Evaluation results show that $MAPE$ <= 10% for 66 to 92% and $MCPE$ <= 15% for 35 to 81% of one-day-ahead predictions with prediction time less than one second. $MAPE$ <= 15% for 71 to 96% of one-week-ahead predictions. HE on-demand scalability characteristic of cloud computing forces cloud service providers to overestimate their resources to meet peak load demand of its customers which happens at different time periods and may not overlap. Due to over-estimation, a large number of cloud resources are idle during off peak hours. Cloud providers also face the problem of allocating resources, keeping in view user"s different job requirements and data center capacity. Different types of users, multiple types of requirements further alleviate the resource allocation problem. Also, demand for cloud resources fluctuate due to today"s usage based pricing plans. In order to manage these demand fluctuations more flexible pricing plans are required to sell resources according to real time market demand. Spot pricing was introduced by Amazon EC2 in December 2009 to minimize operational cost, combat under utilization of its resources and make more profit. Similar to on-demand instances, spot instances offer several instance types comprising different combinations of CPU, memory, storage and networking capacity. Amazon Web Service (AWS) is not the only participant in the spot instance realm. Google Compute Engine launched its preemptible Virtual Machines on September 8, 2015 designed for such type of workloads that can be delayed and are fault tolerant at the same time. Users can bid for spot instances (SIs) where prices are charged at lowest bid price, whereas, pricing on Google Preemptible VMs is fixed at per hour rate. The distinguishing feature of Amazon Elastic Compute Cloud (EC2) spot instance is its dynamic pricing. From customer's perspective, spot instances offer prospects of low cost utility computing at a risk of out-of-bid failure at any time by Amazon EC2. Spot instance reliability depends on the market price and user's maximum bid (limited by their hourly budget). Spot prices vary dynamically with real-time based on demand (user's bid) and supply (resource availability) for spot instance capacity in the data centers across the globe. User's bids for spot instances and control the balance of reliability versus monetary cost. The price for spot instances sometimes can be as low as one eighth of the price of on-demand instances. On the other hand, it is also not uncommon that spot prices surpass on-demand prices in cloud data centers. When the demand is low, spot prices are low because less numbers of users are bidding for the same instance. Therefore, a bidder"s probability of incurring less monetary cost is higher. On the other hand, when the demand is high, users are willing to pay high to get access and hence spot prices increase. Spot pricing in particular is a pricing model targeted for divisible computing jobs that can shift the time of processing to when the computing resources are available at low cost [1]. The primary requirement is that the applications must be time flexible, do not have a steep completion deadline and should be interrupt tolerant. Spot instances are also required for executing certain sudden tasks which do not need reserved instances. The ability to predict spot price lends itself to a variety of applications. Just a few examples of applications suitable to be executed on spot instances are: geospatial analysis, image and video processing, scientific research and computing, data processing, financial modeling and analysis, big data analytics, testing, web crawling and large scale simulations

EXISTING SYSTEM

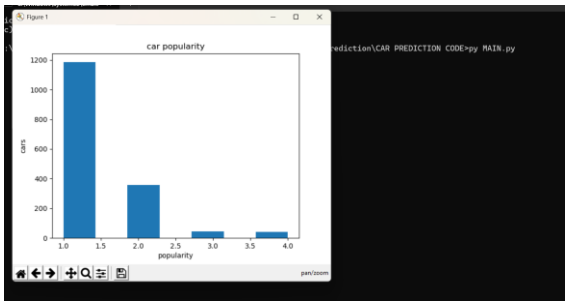
In paper "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks [1]", author has reviewed some classification algorithms such as random forest, gradient boosted trees, artificial neural network and logistic regression to predict 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 463 stocks of the S&P 500. In order to study the predictability of these stocks, author has performed multiples of experiments with these classification algorithms. The obtained result of predicting future prices from the past available data was not up to the mark as the expected result, The author wanted to obtain. However, they successfully showed the vast growth in predictability of European and Asian indexes closed a little while back. In paper "Performance evaluation of predictive models for missing data imputation in weather data [2]", author has suggested a new approach to manage the missing data in weather data by performing various tests with NCDC dataset to assess the prediction error of five methods: linear regression, SVM, random forest, KNN Implementation and kernel ridge. In order to handle the missing values of dataset they performed two actions: 1.removing the entire row which contains missing value and 2. Impute the missing data. They performed both the methods to handle the missing data and compared the observed result. In paper "Amazon EC2 Spot Price Prediction using Regression Random Forests [3]", author has proposed Regression Random Forests (RRFs) model to forecast the Amazon EC2 Spot Price one week ahead and one month ahead. This prediction model would help in planning when to acquire the spot instance, the model also predicts the execution cost and it also suggests the user when to bid in order to minimize the execution cost

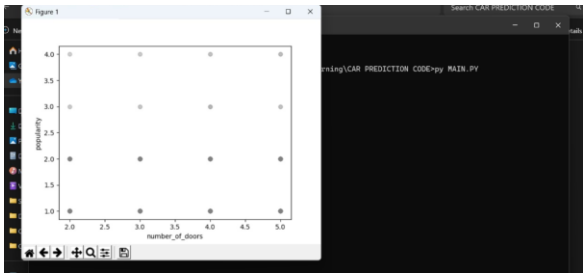
PROPOSED SYSTEM

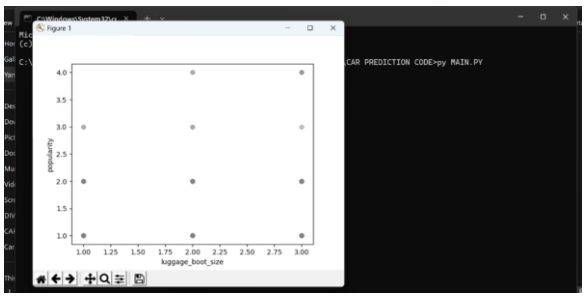
The present system focuses on the introduction of some applicable AI-based strategies that can support existing standard methods of dealing with car popularity. Hence in the present work deep learning strategy is used. As a subset of machine learning. Dl consist of numerous layers of algorithms that provide a different interpretation of the data it feeds on. However, DL is mainly from ML because it Presents data in the system in a different manner. Whereas DL networks works by layers of Artificial Neural Network (ANN), ML algorithms are usually dependent on structured data. Unlike supervised learning which is that the task of learning a function mapping an input to an output on the premise of examples input-output pairs, unsupervised learning is marked by minimum human supervision and will be described as a form of machine learning in search of undetected patterns in an exceedingly data set where no prior

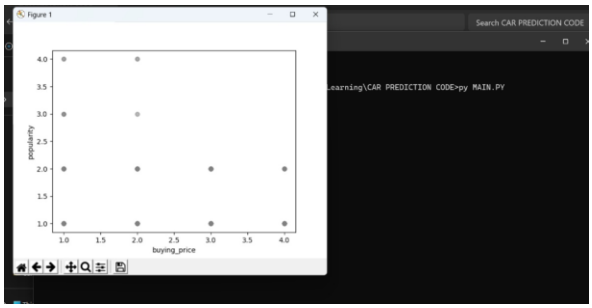
labels exist. DL can be extensively applied for car popularity; however, aims at finding the most effective possible solutions for car popularity related issues. With the aim of foregrounding the enhanced effectiveness of these strategies and techniques, their formation has been informed by. Therefore, this section presents ideas that can enhance and speed up ANN-based methods obtaining process to improve process methods.

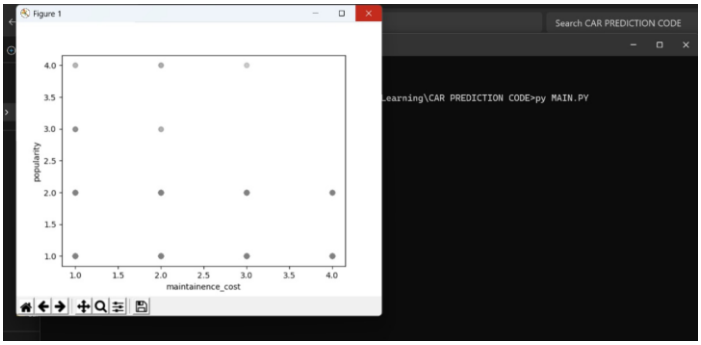
RESULT

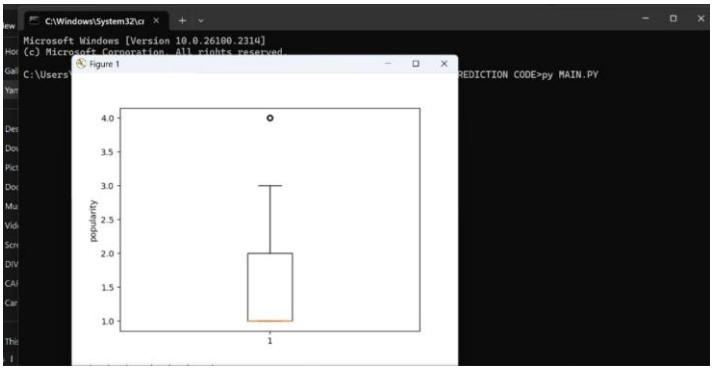


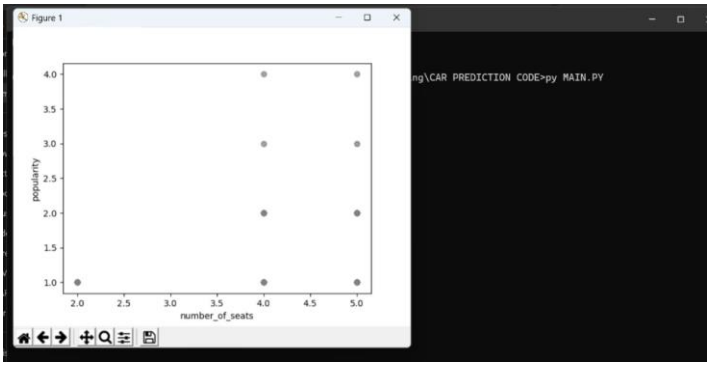


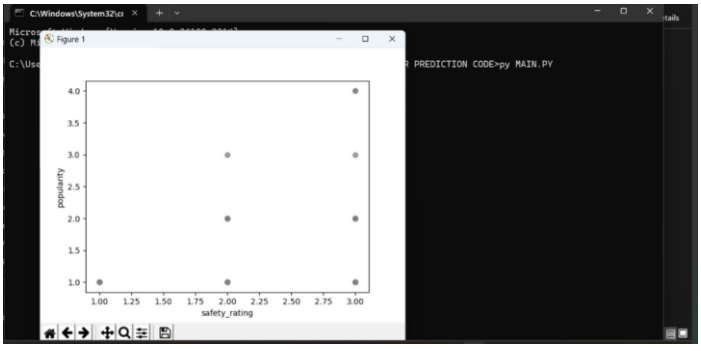












CONCLUSION

Machine Learning is a fast growing approach to solve real world problems. This paper focused on some of the supervised learning algorithms such as Logistic Regression, KNN, SVM and Random Forest for prediction popularity on a scaling measure of [1…4] for a car company. From table 1 it is clear that SVM is giving us the best result. Thus for future work, our focus would be on

modifying SVM model used and will try to make the prediction more accurate. Also implementing the problem using deep learning deep learning and neural network algorithms will be our focus, as they provide more generalization of problems. REFERENCES [1] Jiao, Yang, and Jérémie Jakubowicz. "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks." Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017. [2] Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017. [3] Khandelwal, Veena, Anand Chaturvedi, and Chandra Prakash Gupta. "Amazon EC2 Spot Price Prediction using Regression Random Forests." IEEE Transactions on Cloud Computing, 2017. [4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.. [5] Le, Quoc V., Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y. Ng. "On optimization methods for deep learning." In Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 265-272. Omnipress, 2011. [6] Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005). [7] Olsson, Fredrik. "A literature survey of active machine learning in the context of natural language processing." (2009). [8] Cambria, Erik, and White B. "Jumping NLP curves: A review of natural language processing research." IEEE Computational intelligence magazine 9.2 (2014): 48-57. [9] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24. [10] Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. “A review of machine learning algorithms for text-documents classification.” Journal of advances in information technology, 1(1), pp.4-20. [11] Jiang J. “A literature survey on domain adaptation of statistical classifiers.” URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey. 2008 Mar 6;3. [12] Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. “Reinforcement learning: A survey.” Journal of artificial intelligence research, 4, pp.237-285 [13] Ban, Tao, Ruibin Zhang, Shaoning Pang, Abdolhossein Sarrafzadeh, and Daisuke Inoue. "Referential knn regression for financial time series forecasting." In International Conference on Neural Information Processing, pp. 601-608. Springer, Berlin, Heidelberg, 2013. [14] Dutta, A., Bandopadhyay, G. and Sengupta, S., 2015. “Prediction of stock performance in indian stock market using logistic regression.” International Journal of Business and Information, 7(1). [15] Liaw, A. and Wiener, M. “Classification and regression by randomForest." R news (2002), 2(3), pp.18-22. [16] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. “Random forest: a classification and regression tool for compound classification and QSAR modeling.” Journal of chemical information and computer sciences (2003), 43(6), pp.1947-1958. [17] Smola, A.J. and Schölkopf, B. “A tutorial on support vector regression.” Statistics and computing (2004), 14(3), pp.199-222. [18] Gunn, S.R. “Support vector machines for classification and regression.” ISIS technical report (1998), 14(1), pp.5-16. [19] Williams, N., Zander, S. and Armitage, G. “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification.” ACM SIGCOMM Computer Communication Review (2006), 36(5), pp.5-16. [20] Willmott, C.J. and Matsuura, K. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.” Climate research (2005), 30(1), pp.79- 82

REFERENCES

1. Jiao, Yang, and Jérémie Jakubowicz. "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks." *Big Data (BIG DATA), 2017 IEEE International Conference on.* IEEE, 2017.
2. Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." *Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on.* IEEE, 2017.
3. Khandelwal, Veena, Anand Chaturvedi, and Chandra Prakash Gupta. "Amazon EC2 Spot Price Prediction using Regression Random Forests." *IEEE Transactions on Cloud Computing, 2017.*
4. Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural Networks* 61 (2015): 85-117.
5. Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
6. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
7. Bishop, Christopher M. *Pattern recognition and machine learning.* Springer, 2006.
8. Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction.* MIT press, 2018.
9. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.
10. Chollet, François. *Deep Learning with Python.* Manning Publications, 2018.